AUTOMATIC SUBJECTIVE QUALITY ESTIMATION OF 3D STEREOSCOPIC VIDEOS: NR-RR APPROACH

Hossein Malekmohamadi

School of Computer Science and Informatics
De Montfort University, The Gateway, Leicester, LE1 9BH, UK
Email: hossein.malekmohamadi@dmu.ac.uk

ABSTRACT

A method for estimating subjective quality score of 3D stereoscopic video is proposed which is based on decision trees. The output of this estimation can be fed into encoding and transmission units for compensation. The proposed method operates with minimum dependency on reference video. Content characteristics, no reference (NR) and reduced reference (RR) quality metrics are extracted and summarised prior to training stage. Content features are based on spatio-temporal activities within depth layers. Quality features include NR blockiness, NR blurriness and RR 3D stereoscopic video quality metric. Due to fast and accurate requirements for the quality estimation, decision trees are employed where a 0.94 accuracy is achieved.

Index Terms — Subjective quality, 3D stereoscopic Video, Content characteristics, RR metric, NR metric, Decision Trees.

1. INTRODUCTION

Due to recent advancements of various 3D video applications and services, there is an increasing demand for accurate and operational quality assessment of underlying videos. Accurate estimate for 3D video quality is crucial to video codec development, network protocol planning, quality of service (QoS) monitoring or quality assurance of end-users. Moreover, a component capable of accurate estimation can be combined in video delivery chain in various places like after encoder or before user display. 3D stereoscopic videos are basically attainable in colour plus depth format. They require greater capacities for storage compared to 2D videos and higher transmission rates. Different processes like compression or transmission incur visual artefacts on 3D videos. Concurrently, the purpose in the entire delivery system is to provide end-user with a 3D video that has high quality. Quality evaluations of 3D video can be achieved objectively or subjectively.

Objective metrics measure distortions/qualities incurred during encoding or transmission. They are consistent with time and based on mathematical models to compute distortion. A better understanding of delivered quality is achieved by measuring user's perceived quality through a set of subjective tests. Subjective experiments require real participants which is time consuming and expensive. Unlike objective metrics, subjective evaluations are susceptible to user preferences, equipment or content. Objective metrics can be extracted for each 2D section of 3D stereoscopic video and the relation of metrics to subjective quality can be investigated. However, better estimation of subjective quality requires combination of metrics and preferably having an exclusive 3D quality metric.

Accurate estimatation of subjective quality (ρ) eliminates the need of subjective tests and real participants and help video codec and transmitter to monitor and recover failures. The ultimate goal of many research activities in this area is to bridge this gap to its simplest format: $\rho = f_1(\theta)$, where f_1 is a function of objective quality metrics θ. However, in reality, many influential parameters can be considered as inputs. Parameters like user preferences, aesthetic level for content, price of display technology, eye comfort or ambient light can severely affect human judgement of 3D content. With a fast and accurate estimation of subjective quality, the need of real participants to judge 3D content can be avoided.

Prior research works on relating subjective and objective qualities for 2D and 3D videos have led to some mathematical representations of video quality [1–5]. These representations consider some distortion measures of processed video and compare it to reference video. However, human judgement of video quality based on content characteristics should be considered. This changes the formula to: $\rho = f_2(\theta, \alpha, \phi)$, where ϕ is content characteristics and α represents other factors like human factors or cost. Authors in [6] have used some content characteristics for 2D videos to model subjective quality for 2D videos. In [7], authors suggest a framework to classify content of 2D videos in to 5 classes based on motion estimation features. Content analysis for 2D videos is mainly based on tracking spatial information changes for consecutive frames. In the case of 3D videos, many other factors need consideration. The first element is depth perception. Human eyes distinguish between 3D videos based on amount of depth they have in addition to colour quality. Moreover, 3D motion analysis and 3D texture analysis can help content analysis of 3D video.

For 3D videos, the problem of mathematical representation of perceived quality can be tackled by taking into account content characteristics as well as objective quality metrics. Ideally, this model should be with minimum dependency on reference video, as availability of reference signal is not applicable to receiver side. On the other hand, there is a lack of exclusive objective quality metrics for 3D videos. Several 3D quality metrics can be found in [8–11]. These 3D quality metrics are dependent on availability of reference video. Moreover, they do not consider texture changes during compression and transmission and they are similar to block-based peak signal to noise ratio for each 2D section of 3D video. The last but not the least is depth importance ratio in overall 3D perception needs appropriate exploration. Solving this problem will change the formula to: $\rho = f_3(\hat{\theta}, \alpha, \phi)$, where $\hat{\theta}$ represents reference-free (NR,RR) objective quality metrics. Solving these problems will omit the need of having real human eye observers to assess 3D video content subjectively. To tackle this problem in this paper, the first stage of the proposed technique is content feature extraction. Numbers of depth layers, spatial and temporal features within depth layers are considered to represent video content characteristics. In the second stage, quality features are ex-

tracted. NR Blockiness [12] and NR blurriness [13] are calculated for only colour frames of the 3D stereoscopic video. Furthermore, a RR video quality metric is applied [14]. Content and quality features are used in machine learning algorithms to estimate subjective quality scores.

Decision trees (DT) are selected due to their capabilities for fast and accurate prediction. Extracted features are inputs to DT. Preprocessing of data is performed to summarise 3D stereoscopic video content and quality features which leads to smaller tree size. Summarisation is with non-overlapping windowing for each 5 consecutive frames to reduce the size of input vector by 80%. Learning process of DTs is an inductive process that uses particular facts from data attributes to make more generalised conclusions [15]. After learning process, DT is able to predict the output giving new inputs. There are different methods of learning process of the DTs. In this paper, random subspace method (RSM [16]) and bootstrap aggregating (bagging [17]) are employed. Block diagram of the proposed technique is shown in Figure 1.

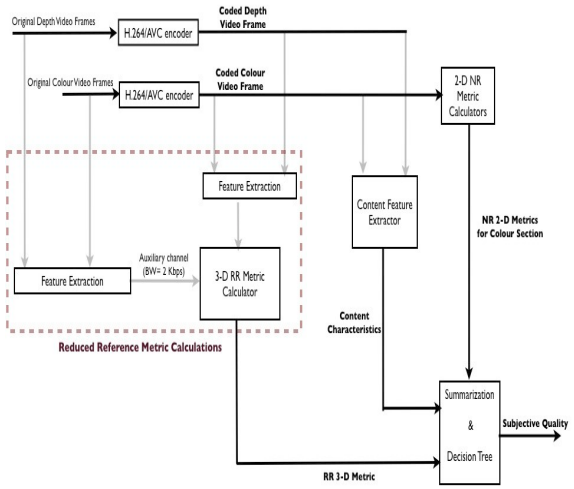

Figure 1. Block diagram of the proposed model.

In this paper, we present method to estimate subjective quality scores for 3D stereoscopic videos in a NR-RR way. This technique extracts content characteristics as well as quality features which are then fed in to the learning algorithms. In Section II, a method for NR-RR content and quality feature extraction is described. Subjective quality estimation is presented in Section III of this paper. Finally, Section IV concludes this paper.

2. NR-RR CONTENT AND QUALITY FEATURE EXTRACTION

One of the important features in 3D stereoscopic videos is their difference in depth perception. Accordingly, finding the number of depth layers is a preliminary stage for content analysis. Depth frame as a grey scale image has 256 possible grey levels ($gl \in$ [0,255]). Afterwards, depth frame is scaled down to 32 grey levels. By scaling, nearby depth layers are combined and considered as one layer. Histogram of scaled depth frame Ω is calculated. Number of peaks in Ω_i (for the i^{th} frame) is the parameter κ_i which is equal to the number of the depth layers for the i^{th} frame of the depth video. Afterwards, stream of peaks of depth video is smoothed and updated to hinder sudden changes from one frame to following frame based on:

$$\hat{\kappa}_i = \begin{cases} \kappa_i, & for\ i = 1, 2(initialisation). \\ \lceil 0.5\kappa_i + 0.5\kappa_{i-1} \rceil, & for\ |\frac{\kappa_i - \kappa_{i-1}}{\kappa_{i-1}}| > 0.5 \\ \lceil \alpha\kappa_i + (1-\alpha)\kappa_{i-1} \rceil, & for\ |\frac{\kappa_i - \kappa_{i-1}}{\kappa_{i-1}}| \leq 0.5 \end{cases} \tag{1}$$

For example, if the i^{th} frame has 7 depth layers and the $i^{\text{th}} - 1$ frame has 3 depth layers, feature extraction algorithm smooths this hard transition and caps the number of depth layers to 5 for the i^{th} frame. Another case is where the i^{th} frame has 4 depth layers and the $i^{\text{th}} - 1$ has 7, feature extraction algorithm smooths this hard transition and caps the number of depth layers to 5 for the i^{th} frame. This approach is similar to fair smoothing. In this paper α is set to 0.97 and the resultant $\hat{\kappa}_i$ is then used for K-means clustering segmentation of corresponding depth frame to generate $\hat{\kappa}_i$ masks. In Figure 2, an example of balanced peaks (depth layers) for a sample 3D videos is shown.

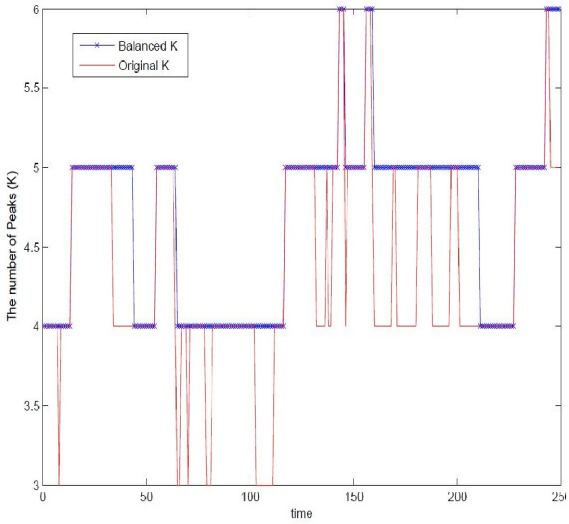

Figure 2. An example of balancing number of depth layers for a sample 3D video (sequence butterfly3 as in [18]).

Each masked colour image has zero and non-zero pixels, where non-zero pixels show the information from the colour frame over the specific part of the segmented depth image. Corresponding colour image is passed through each filter. For each filtered colour frame (the i^{th} frame is I^i in the notations) applying the j^{th} filter $\psi_{i,j}$, spatial feature ξ is defined for the luminance component as:

$$\xi_j^i = \sigma_{\mathbf{V}}(\sigma_{\mathbf{H}}(\hat{S}\ \overbrace{(\psi_{i,j}(I^i)}^{\text{masked colour image}}\ )))_{\substack{i=1,\ldots,N \\ j=1,\ldots,\hat{\kappa}_i}} \tag{2}$$

Where $\hat{S}$ is edge detector, and σ is standard deviation along vertical (**V**) or horizontal (**H**) axis. Another feature is computed from time difference ν, which is computed using luminance components for two consecutive filtered colour frames (the $i^{\text{th}-1}$ and the i^{th} frames) over the correspondent j^{th} filters ($\psi_{i,j}$,$\psi_{i-1,j}$). Time features are computed from:

$$\nu_j^{i,i-1} = \sigma_{\mathbf{V}}(\sigma_{\mathbf{H}}(\psi_{i,j}(I^i) - \psi_{i-1,j}(I^{i-1})))_{\substack{i=1,\ldots,N \\ j=1,\ldots,min(\hat{\kappa}_i,\hat{\kappa}_{i-1})}} \tag{3}$$

Extracted content features can differentiate videos based on spatio-temporal activities within depth layers. 3D Stereoscopic videos have different motion activities in depth layers like high motion in background or low motion in foreground. The next step in the

estimation algorithm is to calculate two NR metrics: blocking and blurriness measures for colour images. Blocking occurs in the block based coding schemes. They appear in block borders as discontinuities or shift in edges along blocks. In this part of the paper, the algorithm introduced in [12] is employed that finds discontinuities along vertical or horizontal lines (borders). Blurriness or luminance bleeding is another artefact caused during compression. This is due to bigger values for QP (hard quantising). This artefact exists along the edges. The method introduced in [13] is used for blurriness. Each 3D stereoscopic video has N (number of frames) blockiness measures and N blurriness measures. To consider overall 3D stereoscopic video quality, the metric in [14] is used. This metric utilises texture information in a content-based combination. Where content characteristics are taken from edge properties. To verify the performance of the 3D stereoscopic video quality metric, it should correspond to opinion scores from conducted subjective experiments.

In this paper, there are 29 reference videos in different types of motion, depth perception, the number of objects, colour palette and moving camera. The spatial resolution for each view (colour/depth) is 960×540 pixels. To have wide range of qualities, different quantisation parameters (QPs) were applied during compression of colour and depth maps. For each 3D video, colour and depth maps were encoded separately using H.264/AVC with variable bitrate coding (VBR) at the QPs: 30, 35, 40, 45, and 50. Baseline profile was used for encoding colour images and depth maps were encoded using high profile. I frames were included every 75 frames for all colour and depth maps (IPPPP...). During coding stage, content adaptive binary arithmetic coding (CABAC) was used. 3D Stereoscopic videos have different spatial information (texture) in depth layer. As an example, difference between a plain background and highly textured background are recognisable through extracted content features. These features and quality features are used in pre-processing before applying to machine learning algorithm to predict subjective quality. Consequently, 725 videos were generated at 25 fps rate. This data set has been already used in [11, 14, 18]. Subjective tests are based on SAMVIQ [19], and user scores are based on continuous scale in the range of [0 (worst)-100 (best)] for overall 3D stereoscopic video quality. In each part of the test, there is an explicit reference and other videos to rate; one video is called hidden reference. Number of participants, the vision test, and ambient light are complying with ITU guidelines for subjective experiments. For each impaired video at least 15 votes were gathered. A 42" inch Philips WOWvx multi-view auto-stereoscopic display with the resolution of 1920×1080 pixels was used for the test. Ambient light and viewing distance were 200 lux and $3m$ respectively.

3. QUALITY ESTIMATION

Each 3D stereoscopic video (N frames) has a continuous opinion score equal to ρ from subjective experiment results:

$$f_3(\lambda_1, \lambda_2, ..., \lambda_N) \simeq \rho \tag{4}$$

After using non-overlapping window with size 5, each 3D stereoscopic video is summarised with a set of group of frames (Λ) belonging to the same quality score region as the whole video:

$$f_3(\Lambda_{1,k}, \Lambda_{2,k}, ..., \Lambda_{N,k}) \simeq \rho \tag{5}$$

where $\Lambda_{j,k} = G(\lambda_j - k, ..., \lambda_j, ..., \lambda_j + k)$ is the grouping function. In this paper, content and quality features are summarised with non-overlapping windowing as grouping function. Summarised features are considered as input vector which are fed to different learning algorithms. A non-overlapping windowing for 5 frames is used for summarisation that gives 80% decrease of input vector size for our datasets. Non-overlapping windowing can be treated as a time buffer for duration of 200 ms for videos with 25 fps. Accordingly, input vector consists of content-quality features which are spatial information within depth layers as shown in [5], time information within depth layers as shown in [6], NR blockiness for colour video frames as shown in [12], NR blurriness for colour video frames as shown in [13] and RR 3D stereoscopic video quality metric as shown in [14]. For the metric in [14], a ratio of 0.8 for colour and 0.2 for depth is applied in this paper. Decision trees are trained and tested with the summarised features. Random subspace method [16], is an ensemble classifier and a generalization of the random forests algorithm. Bagging [17], is an ensemble algorithm to improve stability and accuracy of classification. It helps to avoid over-fitting. For 725 impaired videos, summarisation gives 725×(250/5)=36250 instances per feature. As an example, totally, we have 36250 measures of blockiness for all our data set. Half of the instances are kept for training. Number of instances for training and test sets is 18125. Train and test sets are selected randomly among all data set where GOFs (Λ) for each impaired 3D stereoscopic video are in either training or test set. In other words, train and test sets have proper distance. Test results for two learning algorithms are shown in Table 1.

Table 1. Experimental Results for the learning algorithms: Tree size (S) is the number of nodes, CC is the correlation coefficients, and RMSE is root-mean-square error.

Algorithm	S	CC	RMSE
Random Subspace	2205	0.94	6.551
Bagging	3173	0.934	7.625

It is evident that DT with random subspace algorithm performs better with 94% accuracy and remarkably reduced tree size (number of nodes in DT). This method modifies training data in the feature space. This method is fast and accurate to apply in to frame-by-frame (or bufferred) estimation of subjective quality. Future changes to this method will include bit-rate considerations. To make the system completely NR, random subspace algorithm is trained and tested leaving out the 3D RR metric introduced in ([14]). Results of this investigation are shown in Table 2. Omitting the RR metric leads to 0.22 decrease in estimation accuracy, 10 increase in RMSE, and 57.14% decrease in tree size, which means many nodes in DT are incorporating the proposed metric.

Table 2. Experimental Results for RSM learning algorithm with only NR features: Tree size (S) is the number of nodes, CC is the correlation coefficients, and RMSE is root-mean-square error.

Algorithm	S	CC	RMSE
Random Subspace	945	0.7208	16.659

As shown in results, NR-RR estimation of subjective quality in this paper has good accuracy. Furthermore, due to nature of training algorithms used in this paper, implementations of the proposed estimation are lightweight in terms of computational complexity and memory. This system estimates subjective quality score at the end of each buffer time which is 200 ms. There are several ways to improve this estimation algorithm. One is to deploy bit-rate considerations in training algorithms and involve them a new feature set. Some improvements can be achieved by adding other NR-RR quality features that consider other perspective of visual quality. However, both proposed remedies will increase tree size

that is a very important factor. The concept behind this system is extensible to situations where subjective score is gathered in a real-time fashion. It means that users rate video quality frequently in smaller fractions of time rather than at the end of 8s or 10s, which are current standards. For this purpose, there is a high demand for comprehensive subjective quality assessment standards that are capable of this.

4. CONCLUSION

A method for fast and accurate subjective quality estimation of 3D stereoscopic video is proposed. The first stage of the model extracts content and quality features with minimum dependency on reference video. Furthermore, with the help of non-overlapping windowing, the number of the content-quality features is reduced significantly preserving the important information concurrently. Summarised features are fed into a decision tree to predict opinion score inductively. Maximum measure of success of the proposed system for test sequences is 0.94. The output of the proposed system can be fed into other units for compensation to prevent further degradation of perceived quality.

5. REFERENCES

[1] Z. W. Ligang, Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment using structural distortion measurement," in *in Proc. IEEE Int. Conf. Image Proc*. Citeseer, 2002.

[2] S. Winkler, "Video quality measurement standardscurrent status and trends," in *Information, Communications and Signal Processing, 2009. ICICS 2009. 7th International Conference on*. IEEE, 2009, pp. 1–5.

[3] C. T. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 304–318, 2009.

[4] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Electronic Imaging: Science & Technology*. International Society for Optics and Photonics, 1996, pp. 450–461.

[5] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[6] M. Ries, C. Crespi, O. Nemethova, and M. Rupp, "Content based video quality estimation for h. 264/avc video streaming," in *Wireless Communications and Networking Conference, 2007. WCNC 2007. IEEE*. IEEE, 2007, pp. 2668–2673.

[7] A. Khan, L. Sun, and E. Ifeachor, "Video quality assessment as impacted by video content over wireless networks," *International Journal on Advances in Networks and Services*, vol. 2, no. 2&3, 2009.

[8] C. T. Hewage and M. G. Martini, "Edge-based reduced-reference quality metric for 3-d video compression and transmission," *IEEE Journal of selected topics in signal processing*, vol. 6, no. 5, pp. 471–482, 2012.

[9] P. Joveluro, H. Malekmohamadi, W. C. Fernando, and A. Kondoz, "Perceptual video quality metric for 3d video quality assessment," in *3DTV-conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2010*. IEEE, 2010, pp. 1–4.

[10] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B. Akar, "Towards compound stereo-video quality metric: a specific encoder-based framework," in *Image Analysis and Interpretation, 2006 IEEE Southwest Symposium on*. IEEE, 2006, pp. 218–222.

[11] G. Nur, H. K. Arachchi, S. Dogan, and A. Kondoz, "Extended vqm model for predicting 3d video quality considering ambient illumination context," in *3DTV Conference: The True Vision-Capture, Transmission and Display of 3D Video (3DTV-CON), 2011*. IEEE, 2011, pp. 1–4.

[12] Z. Wang, A. C. Bovik, and B. Evan, "Blind measurement of blocking artifacts in images," in *Image Processing, 2000. Proceedings. 2000 International Conference on*, vol. 3. Ieee, 2000, pp. 981–984.

[13] P. Marziliano, F. Dufaux, S. Winkler, and T. Ebrahimi, "A no-reference perceptual blur metric," in *Image Processing. 2002. Proceedings. 2002 International Conference on*, vol. 3. IEEE, 2002, pp. III–III.

[14] H. Malekmohamadi, A. Fernando, and A. Kondoz, "A new reduced reference metric for color plus depth 3d video," *Journal of Visual Communication and Image Representation*, vol. 25, no. 3, pp. 534 – 541, 2014, qoE in 2D/3D Video Systems. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S104732031300223X

[15] I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2016.

[16] T. K. Ho, "The random subspace method for constructing decision forests," *IEEE transactions on pattern analysis and machine intelligence*, vol. 20, no. 8, pp. 832–844, 1998.

[17] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[18] H. Malekmohamadi, W. Fernando, and A. Kondoz, "Content-based subjective quality prediction in stereoscopic videos with machine learning," *Electronics letters*, vol. 48, no. 21, pp. 1344–1346, 2012.

[19] J.-L. Blin, "Samviq–subjective assessment methodology for video quality," *Rapport technique BPN*, vol. 56, p. 24, 2003.